programa final

Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais

INFORMAÇÃO PARA UMA SOCIEDADE MAIS JUSTA

III Conferência Nacional de Geografia e Cartografia

IV Conferência Nacional de Estatística

Reunião de Instituições Produtoras
Fórum de Usuários
Seminário "Desafios para Repensar o Trabalho"
Simpósio de Inovações
Jornada de Cursos
Mostra de Tecnologias de Informação

27 a 31 de maio de 1996
Rio de Janeiro, RJ BRASIL

IBGE 60 anos
1936-1996

Uma das maneiras de olhar o ofício de produzir informações sociais, econômicas e territoriais é como arte de descrever o mundo. Estatísticas e mapas transportam os fenômenos da realidade para escalas apropriadas à perspectiva de nossa visão humana e nos permitem pensar e agir à distância, construindo avenidas de mão dupla que juntam o mundo e suas imagens. Maior o poder de síntese dessas representações, combinando, com precisão, elementos dispersos e heterogêneos do cotidiano, maior o nosso conhecimento e a nossa capacidade de compreender e transformar a realidade.

Visto como arte, o ofício de produzir essas informações reflete a cultura de um País e de sua época, como essa cultura vê o mundo e o torna visível, redefinindo o que vê e o que há para se ver.

No cenário de contínua inovação tecnológica e mudança de culturas da sociedade contemporânea, as novas tecnologias de informação - reunindo computadores, telecomunicações e redes de informação - aceleram aquele movimento de mobilização do mundo real. Aumenta a velocidade da acumulação de informação e são ampliados seus requisitos de atualização, formato - mais flexível, personalizado e interativo - e, principalmente, de acessibilidade. A plataforma digital vem se consolidando como o meio mais simples, barato e poderoso para tratar a informação, tornando possíveis novos produtos e serviços e conquistando novos usuários.

Acreditamos ser o ambiente de conversa e controvérsia e de troca entre as diferentes disciplinas, nas mesas redondas e sessões temáticas das Conferências Nacionais de Geografia, Cartografia e Estatística e do Simpósio de Inovações, aquele que melhor enseja o aprimoramento do consenso sobre os fenômenos a serem mensurados para retratar a sociedade, a economia e o território nacional e sobre as prioridades e formatos das informações necessárias para o fortalecimento da cidadania, a definição de políticas públicas e a gestão político - administrativa do País, e para criar uma sociedade mais justa.

Simon Schwartzman
Coordenador Geral do ENCONTRO

Promoção

Fundação Instituto Brasileiro de Geografia e Estatística
IBGE
Fundação Instituto Brasileiro de Geografia e Estatistica
IBGE
Associação Brasileira de Estudos Populacionais
ABEP

Co-Promoção

Associação Brasileira de Estatística
ABE
Associação Brasileira de Estudos do Trabalho
ABET
Associação Brasileira de Pós-graduação em Saúde Coletiva
ABRASCO
Associação Nacional de Centros de Pós-graduação em Economia
ANPEC
Associação Nacional de Pós-graduação e Pesquisa em Ciências Sociais
ANPOCS
Associação Nacional de Pós-graduação e Pesquisa em Geografia
ANPEGE
Associação Nacional de Pós-graduação e Pesquisa em Planejamento Urbano e Regional
ANPUR
Sociedade Brasileira de Cartografia
SBC

Apoio

Federação das Indústrias do Estado do Rio de Janeiro
FIRJAN
Academia Brasileira de Letras
ABL
Conselho Nacional de Pesquisas
CNPq
Financiadora de Estudos e Projetos
FINEP
Revista Ciência Hoje

Companhia do Desenvolvimento do Planalto Central
CODEPLAN (DF)
Empresa Metropolitana de Planejamento da Grande São Paulo S/A
EMPLASA (SP)
Empresa Municipal de Informática e Planejamento S/A
IPLANRIO (RJ)
Fundação Centro de Informações e Dados do Rio de Janeiro
CIDE (RJ)
Fundação de Economia e Estatística
FEE (RS)
Fundação de Planejamento Metropolitano e Regional
METROPLAN (RS)
Fundação Instituto de Planejamento do Ceará
IPLANCE (CE)
Fundação João Pinheiro
FJP (MG)
Fundação Joaquim Nabuco
FUNDAJ (PE)
Fundação Sistema Estadual de Análise de Dados
SEADE (SP)
Instituto Ambiental do Paraná
IAP (PR)
Instituto de Geociências Aplicadas
IGA (MG)
Instituto de Pesquisas Econômicas, Administrativas e Contábeis
IPEAD (MG)
Instituto do Desenvolvimento Econômico Social do Pará
IDESP (PA)
Instituto Geográfico e Cartográfico
IGC (SP)
Instituto de Apoio à Pesquisa e ao Desenvolvimento "Jones dos Santos Neves"
IJSN (ES)
Instituto Paranaense de Desenvolvimento Econômico e Social
IPARDES (PR)
Processamento de Dados do Município de Belo Horizonte S/A
PRODABEL (MG)
Superintendência de Estudos Econômicos e Sociais da Bahia
SEI (BA)

Organização

Coordenação Geral

Simon Schwartzman

Comissões de Programa

Confege

César Ajara (IBGE)
Denizar Blitzkow (USP)
Jorge Marques (UFRJ)
Lia Osório Machado (UFRJ)
Mauro Pereira de Mello (IBGE)
Speridião Faissol (UERJ)
Trento Natali Filho (IBGE)

Confest

José A. M. de Carvalho (UFMG)
José Márcio Camargo (PUC)
Lenildo Fernandes Silva (IBGE)
Teresa Cristina N. Araújo (IBGE)
Vilmar Faria (CEBRAP)
Wilton Bussab (FGV)

Comissão Organizadora

Secretaria Executiva - Luisa Maria La Croix
Secretaria Geral - Luciana Kanham
Confege, Confest e Simpósio de Inovações
Anna Lucia Barreto de Freitas, Evangelina X.G. de Oliveira,
Jaime Franklin Vidal Araújo, Lilibeth Cardozo R.Ferreira e
Maria Letícia Duarte Warner
Jornada de Cursos - Carmen Feijó
Finanças - Marise Maria Ferreira
Comunicação Social - Micheline Christophe e Carlos Vieira
Programação Visual - Aldo Victorio Filho e
Luiz Gonzaga C. dos Santos
Infra-Estrutura - Maria Helena Neves Pereira de Souza
Atendimento aos Participantes - Cristina Lins
Apoio
Andrea de Carvalho F. Rodrigues, Carlos Alberto dos Santos,
Delfim Teixeira, Evilmerodac D. da Silva, Gilberto Scheid,
Héctor O. Pravaz, Ivan P. Jordão Junior,
José Augusto dos Santos, Julio da Silva, Katia V. Cavalcanti, Lecy Delfim,
Maria Helena de M. Castro, Regina T. Fonseca,
Rita de Cassia Ataualpa Silva e Taisa Sawczuk

Registramos ainda a colaboração de técnicos das diferentes áreas do IBGE, com seu trabalho, críticas e sugestões para a consolidação do projeto do ENCONTRO.

SOME METHODOLOGICAL ISSUES IN STATISTICAL DISCLOSURE CONTROL

L.C.R.J. Willenborg, A.G. de Waal and W.J. Keller[1]

ABSTRACT

In the last decade the demand for detailed information has increased considerably. The increased demand for detailed information becomes clear from the data that are released by statistical office. Whereas in the old days relatively small two-ways tables were sufficient to satisfy most of the users' demands, nowadays large three- and higher-dimensional tables are no longer an exception. Microdata sets, i.e. data sets containing data on individual respondents, are relatively new products of statistical offices. Such a microdata set contains a wealth of information. However, both the release of large tables and of microdata sets lead to considerable problems when trying to protect the privacy of respondents. In this paper we examine how Statistics Netherlands is dealing with the problems of statistical disclosure control. The emphasis is on disclosure control for microdata sets rather than for tables, because disclosure control for microdata is a relatively new, and still controversial, subject. The current rules and techniques for microdata sets that are applied at Statistics Netherlands are examined. Applying these rules and techniques is not a straightforward matter. A number of methodological problems must be solved in order to apply these rules and techniques appropriately. Moreover, a number of potential improvements for our rules are examined. All these potential improvements require further theoretical research, however. Finally, a number of similarities and differences between statistical disclosure control for microdata and for tables is pointed out.

Keywords: statistical disclosure control, microdata, re-identification

1. INTRODUCTION

In the last decade the demand for detailed information has increased considerably. This is mainly due to the great power of modern PC's, which enables researchers to analyse large data sets by themselves, whereas in former days only the statistical offices were able to analyse such data sets. The increased demand for detailed information becomes clear from the data that are released by statistical offices. Whereas in the old days relatively small two-way tables were sufficient to satisfy most of the users' demands, nowadays large three- and higher-dimensional tables are no longer an exception. Microdata sets, i.e. data sets containing data on individual respondents, are relatively new products of statistical offices. Such a microdata set contains a wealth of information. However, both the release of large tables and of microdata sets lead to considerable problems when trying to protect the privacy of respondents. Especially microdata sets create interesting challenges in the field of statistical disclosure control (SDC).

[1] Statistics Netherlands, Division Research and Development, PO Box 4000, 2270 JM Voorburg, The Netherlands

In this paper we examine how Statistics Netherlands is dealing with the problems of SDC. The emphasis is on SDC for microdata sets rather than for tables, because SDC for microdata is a relatively new, and still developing, subject. Statistics Netherlands releases two kinds of microdata sets, namely public use files and microdata sets for research. The current SDC-rules and techniques for these microdata sets that are applied at Statistics Netherlands are examined in more detail in Section 2.

Applying these rules and techniques is not a straightforward matter. On the contrary, a number of methodological problems must be solved in order to apply these rules and techniques appropriately. For instance, the population frequency of a combination of values of identifying variables is usually not known, and has to estimated from a sample. Another important problem is how to apply the SDC-techniques in such a way that the resulting microdata set is considered safe, while the information loss due to these measures is minimised. These and other subjects are examined in Section 3.

Although Statistics Netherlands has SDC-rules for its microdata sets, this does not imply that they will remain fixed forever. In the future these rules and techniques are likely to change. This will occur not only because society itself is constantly changing, but also because of some dissatisfaction with the present rules. We can imagine better rules, but these rules, unfortunately, require a methodological understanding that is greater than our present understanding. In Section 4, a number of potential improvements for our rules are examined. All these potential improvements require further theoretical research, however.

Of course, apart from microdata sets Statistics Netherlands also releases tables. Several SDC-problems for tables are discussed in Section 5. The aim of that discussion is not to provide a complete description of the way in which Statistics Netherlands handles SDC-problems for tables, but rather to illustrate a number of similarities and differences between SDC for microdata and for tables. The paper is concluded with a brief discussion in Section 6.

2. SDC FOR MICRODATA AT STATISTICS NETHERLANDS

The basic idea of the SDC-rule for microdata applied at Statistics Netherlands is that certain combinations of values of variables should occur frequently enough in the population. We begin our exposition by explaining the underlying philosophy of this basic idea.

2.1 The basic idea

The aim of statistical disclosure control is to limit the risk that sensitive information of individual respondents can be disclosed from a data set. Such a disclosure of sensitive information of an individual respondent can occur after this respondent has been re-identified, i.e. after it has been deduced which record in a given microdata set corresponds to this particular individual. SDC should therefore hamper the re-identification of individual respondents. Re-identification can take place when several values of so-called identifying variables, i.e. variables of which the value can be used alone or in combination with the values of other such variables to re-identify a respondent, are taken into consideration. The values of these identifying variables can be assumed known to friends and acquaintances of a respondent. Examples of identifying variables are 'Place of residence', 'Sex' and 'Occupation'.

An important concept in the theory of re-identification is a *key*, i.e. a combination of identifying variables. The dimension of a key is the number of identifying variables present in this key. Re-identification of a respondent can occur when this respondent is unique in the population with respect to a certain key value, i.e. a combination of values of identifying variables. Hence, uniqueness of respondents in the population with respect to certain key values should be avoided. When a respondent appears to be unique in the population with respect to a particular key value, then disclosure control measures might be called for to protect this respondent against re-identification. In practice, however, it is a bad idea to try to prevent only the occurrence of respondents in the data file who are unique in the population (with respect to a certain key). For this several reasons can be given. Firstly, there is a practical reason: unicity in the population, in contrast to unicity in the data file, is hard to establish. There is no way to determine whether a person who is unique in the data file (with respect to a certain key) is also unique in the population. Secondly, an intruder may use another key than the key(s) considered by the data protector. For instance, the data protector may consider only keys consisting of at most three variables while the intruder may use a key consisting of four variables. Therefore, it is better to try to avoid the occurrence of combinations of scores in the data file that are rare in the population instead of trying to avoid only population-uniques in the data file.

When the frequency of a combination of values of identifying variables is sufficiently high then this combination is considered safe, otherwise it is considered unsafe. If a record contains any unsafe combinations then this record may not be published in its present form, and appropriate SDC-measures should be applied.

2.2 Public use files and microdata for research

Now that we have examined one of the basic ideas behind the SDC-rules applied at Statistics Netherlands it is time to consider the two kinds of microdata sets and their corresponding rules that are disseminated by Statistics Netherlands. In neither kind of microdata sets, formal identifiers, such as name or address, are published, of course. The first kind of microdata sets are so-called *public use files*. A public use file can be obtained by everybody. The data contained in a public use file should be at least one year old. The keys that have to be examined for a public use file are all combinations of two identifying variables. The (estimated) population frequency of a value of an identifying variable has to be at least d_1, the (estimated) population frequency of a bivariate combination has to be at least d_2, where d_1 and d_2 are fixed threshold values ($d_1>d_2$). The number of identifying variables is limited, and identifying variables referring directly to a region of residence, work or education, such as 'Place of residence' are not included in a public use file. Very sensitive variables, such as variables on sexual behaviour or criminal activities, may not be included in a public use file. Sampling weights have to be examined before they can be included in a public use file, because there are situations in which these weights can give additional identifying information. For instance, when a certain subpopulation is oversampled then this subpopulation can be recognised by the relatively low weights associated with its members in the sample. Sampling weights may only be published when they do not provide additional information that can be used for disclosure purposes. Re-grouping of households should be prevented, because households are more likely to be unique on a low-dimensional key than the constituting individuals,. For this we check whether combinations of so-called household variables, i.e. variables relating directly to a household such as 'Number of persons in the household' and 'Occupation of the head of the household', occur in sufficiently many different households. When this is not the case SDC-measures, such as recoding the household variables, should be taken. Re-grouping of the records by taking into account (by an intruder) that the records in a microdata set are usually sorted in a particular order, e.g. all members of a households are placed consecutively or all respondents from the same region are placed consecutively, should also be prevented. Such a re-grouping of the records could provide additional identifying information. To prevent re-grouping the rules demand that before a public use file is released the order of the records should be randomised. Finally, the rules hardly allow information on region of residence, work or education in a public use file. Only one of these three kinds of regional information may be included in a public use file. Moreover, it has to be checked whether the regions that can be distinguished in the microdata set are sufficiently scattered over the country. When these regions would not be sufficiently scattered they would form compact areas in which it would be relatively easy to trace a particular respondent. Checking whether the regions that can be distinguished in the microdata set are sufficiently scattered over the country involves all variables that provide any regional information. For example, the variable 'Number of swimming pools in your place of residence' is taken into consideration while performing the checks.

The second kind of microdata sets are so-called *microdata sets for research*. A microdata set for research can only be obtained by well-respected (statistical) research offices. The information content of a microdata set for research is much higher than that of a public use file. The number of identifying variables is not limited and identifying variables with much regional detail, such as 'Place of residence', may be included in a microdata set for research. Because of the high information content of a microdata set for research, researchers have to sign a declaration stating that they will protect any information about an individual respondent that might be disclosed by them. Detailed regional information may be included in a microdata set for research. The variables with information on region of residence, work and education should be crossed. The resulting variable is called *the regional variable*.

The keys that have to be examined for a microdata set for research include three-way combinations of the regional variable with variables describing the sex, ethnic group or nationality of a respondent with an ordinary identifying variable. The (estimated) population frequency of these trivariate combinations should be at least d_0, where d_0 is a fixed threshold value. The value of d_0 is less than the threshold value d_2 for bivariate combinations in the case of a public use file. The number of persons in a region that can be distinguished in a microdata set for research should be at least 10,000. Finally, the information that may be given on 'Occupation', 'Employer' and 'Education' of a respondent depends on how much regional information is given in this data set. If much regional information is given then little information may be provided on the other three subjects. For more information on the kinds of microdata sets released by Statistics Netherlands and their rules we refer to Keller and Willenborg (1993).

2.3 Disclosure protection measures

The SDC-rules of a statistical office to determine whether a microdata set is considered safe for release is an important part of its SDC-policy. Another, equally important, part is formed by the techniques that are applied when it turns out that a particular microdata set is considered unsafe for release. Statistics Netherlands advocates two SDC-techniques to protect unsafe combinations in microdata sets, namely global recoding and local suppression. In case of global recoding several categories of a variable are collapsed into a single one. A global recoding is applied to the whole data set, not only to the unsafe part of the set. This is done to obtain an uniform categorisation of each variable. When local suppression is applied one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. A local suppression is only applied to a particular value. Both global recoding and local suppression lead to a loss of information, because either less detailed information is provided or some information is not given at all. Note that local suppressions may lead to a bias in the data, because the extreme combinations are removed from the data set. This is not considered to be a serious problem by Statistics Netherlands because we try to limit the number of local suppressions as much as possible, i.e. we try to protect most unsafe combinations by global recodings. Except that much information would be lost when certain unsafe combinations were protected by global recodings only we apply local suppressions. An optimal mix of global recodings and local suppressions has to be found in order to make the information loss due to these SDC-measures as low as possible.

The above two techniques are non-perturbative ones, i.e. they do not modify the values of the variables. There are two main reasons why Statistics Netherlands does not apply any perturbative techniques at the moment. Firstly, it seems hard to ensure the statistical quality of the resulting data when perturbations would be applied on a large scale. However, when the number of values that are perturbed is small, i.e. when the perturbations would be applied as parsimoniously as the local suppressions are, then the statistical quality of the resulting microdata set is bound to be acceptable (Cf. Barnes, 1995). Secondly, a disadvantage of perturbation is that the resulting data may be impossible or very unlikely. For example, when the variable age has been perturbed there may be a record in the resulting data set with the following values 'Age=12 years' and 'Marital status=widowed'. This is highly implausible.

Checking all the SDC-rules is a time-consuming and error prone job. Therefore Statistics Netherlands is developing a general purpose software package for SDC of microdata (Cf. De Jong, 1992; De Waal and Willenborg, 1994b; Van Gelderen, 1995; Pieters and De Waal, 1995; De Waal and Pieters, 1995). The package, ARGUS, should enable the statistical office to analyse the data and to carry out suitable protection measures. The structure of the package is such that it will be possible to specify different disclosure control rules. This implies that the package will be suited for other statistical offices too. Moreover, it should be possible to incorporate changes in the rules fairly easily in the package. The strongest feature of the current version of the package is the possibility to determine the necessary local suppressions automatically and optimally, i.e. the number of local suppressions is minimised, after the global recodings have been determined interactively.

The further development of ARGUS is a major goal of an ESPRIT-project on SDC. In fact, a package for the SDC of microdata, μ-ARGUS, and a package for the SDC of tabular data, τ-ARGUS will be developed. The participating institutions in this project are Eindhoven University of Technology, the University of Manchester, the University of Leeds, the Office of Population Censuses and Surveys (OPCS), the Istituto Nazionale di Statistica (ISTAT), the Consortio Padova Ricerche (CPR), and Statistics Netherlands, acting as project co-ordinator.

3. METHODOLOGICAL PROBLEMS

As one can conclude from Section 2, rules for SDC of microdata applied at Statistics Netherlands are for a substantial part based on testing whether scores on certain keys occur frequently enough in the population. A few problems arising here are the determination of the keys that have to be examined, the way to estimate the number of persons in the population that score on a certain key, and how to determine appropriate SDC-measures.

The keys that have to be examined at Statistics Netherlands are prescribed by the rules once it has been determined which variables should be considered identifying. For this, Statistics Netherlands applies a number of criteria, such as the visibility of the values of a variable and the tractability of these values. These criteria do not reach a definite verdict for all variables, however. In many cases deciding whether a variable should be considered identifying is a matter of personal judgement.

When applying one of the threshold rules mentioned in Section 2 to determine whether or not a combination of values of identifying variables occurs frequently enough in the population we are generally posed with the problem that we do not know this population frequency. Often we only have a sample available to us to estimate the population frequency. Estimating the population frequency of a certain combination is especially a problem when one of the values corresponds to a region. In fact, we then have to estimate the population frequency of a combination in a certain region instead of in the entire country.

For large regions it is possible to use an interval estimator to test whether or not a key value occurs frequently enough. This interval estimator is based on the assumption that the number of times that a key value occurs in the population is Poisson distributed (Cf. Pannekoek, 1995). However, for relatively small regions the number of respondents is low, which causes this estimator to have a high variance which in turn leads to a lot of records that need to be modified. To estimate the number of times that a key value occurs in a small region we therefore suggest to apply a point estimator.

A simple point estimator for the number of times that a certain key value occurs in a region is the direct point estimator. The fraction of a key value in a region i is estimated by the sample frequency of this key value in region i divided by the number of respondents in region i. The population frequency is then estimated by this estimated fraction multiplied by the number of inhabitants in region i. When the number of respondents in region i is low, which is often the case, the direct estimator is unreliable.

Another point estimator is based on the assumption that the persons who score on a certain key value are distributed homogeneously over the population. In this case the fraction of a key value in region i can be estimated by the fraction in the entire sample. The advantage of this, so-called, synthetic, estimator is that the variance is much smaller than the variance of the direct estimator. Unfortunately, the homogeneity assumption is usually not satisfied which causes the estimator to be biased. However, a combined estimator can be constructed with both an acceptable variance and an acceptable bias by combining this estimator and the direct estimator. Such a combined estimator has been tested in Pannekoek and De Waal (1995) and the results are encouraging.

Another practical problem that deserves attention is top-coding of extreme values of continuous (sensitive) variables. These extreme values may lead to re-identification because these values are rare in the population. At the moment we at Statistics Netherlands use an interval estimator to test whether there is a sufficient number of individuals in the population who score on a 'comparable' value of the continuous variable (Cf. Pannekoek, 1992), although we may apply a point estimator in the future. If there is a sufficient number of persons in the population that score on a comparable value, then the extreme value may be published, otherwise the extreme value must be locally suppressed or the corresponding variable should be globally recoded. In order to apply this method in practice it remains to specify what is meant by 'sufficient' and by 'comparable'.

Some important practical problems occur when determining which protection measures should be taken when a microdata set appears to be unsafe. In that case the original data set must be modified in such a way that the information loss due to SDC-measures is as low as possible while the resultant data set is considered safe. As has been mentioned in Section 2

Statistics Netherlands currently applies only local suppressions and global recodings to protect microdata sets. Our aim is therefore to determine the optimal mix of these local suppressions and global recodings.

In De Waal and Willenborg (1994a) 0-1 integer programming formulations for determining the optimal local suppressions are presented. These formulations all aim to minimise the information loss while protecting the microdata set. They differ with respect to the way in which a data protector wants to measure this information loss. For example, the data protector can decide to minimise the total number of locally suppressed values, or he can decide to minimise the number of different locally suppressed categories. The data protector can also decide to combine these goals, e.g. minimise the number of different locally suppressed categories given that the total number of locally suppressed values has been minimised. As will be clear the information loss due to local suppressions only can be determined by very simple measures, namely the total number of locally suppressed values or the number of different locally suppressed categories.

Determining the optimal global recodings, or an optimal mix of global recodings and local suppressions is much more difficult. Measuring the information loss due to global recodings is already a problem. A simple information measure is not available, in contrast to the case of local suppressions only. In De Waal and Willenborg (1995c) this problem of measuring the information loss is solved by using the entropy. Both the information loss due global recodings and the information loss due to local suppressions can be evaluated by this measure. To evaluate the information measure based on the entropy it is necessary to specify a model for the way in which the users of the microdata will deal with the missing values. For instance, we can assume that for a quantitative variable they will simply replace each missing value by the average value of this variable. In this case the information loss due to local suppression will be rather high. When the users of the data are supposed to be somewhat smarter, we can assume that they use multivariate techniques to impute the missing values. In other words, we could assume that the users of the data explicitly take the logical and statistical dependencies of the data into account. In this case the actual information loss due to local suppression will be less high. Note that if the information loss due to local suppressions would be very low, these local suppressions are less effective for protecting data than they appear to be at first sight (also see the remarks on complex microdata in Section 4).

For public use files a number of additional problems have to be solved. For instance, for these files sampling weights may not provide additional identifying information. In De Waal and Willenborg (1995a) it is shown however that in many cases such additional identifying information can be obtained from the sampling weights. There are two ways to prevent this derivation of additional identifying information, namely subsampling and adding noise to the sampling weights. Subsampling, i.e. deleting records from the microdata set, has the advantage that it is easy to apply, but has the disadvantage that it may lead to a considerable loss of information. Adding noise is more difficult to apply. On the one hand it should not be possible to derive additional identifying information from the sampling weights after noise has been added, on the other hand the statistical quality of the sampling weights should be sufficiently high. The former condition can easily be satisfied by adding much noise to the sampling weights, but then the latter condition will be violated, and vice versa.

4. TOWARDS A FOUNDATION OF SDC FOR MICRODATA

The SDC-rules and techniques described in the previous sections are based on intuitive reasoning rather than on a formal mathematical model. All rules and techniques reduce the re-identification risk, but it is not possible to evaluate this reduction of the re-identification risk. This is a somewhat undesirable situation. Ideally, we would like to have a model for the re-identification risk per record. When such a model would be available the SDC-rules would only have to prescribe the keys that should be checked and the maximum risk that the statistical office that is releasing a particular microdata set is willing to take. When the actual re-identification risk of a record is less than this maximum risk then the record may be published without modifications, otherwise the record should be modified.

Several efforts have been made to develop such a re-identification risk per record model. Some of these efforts did not take the 'noise' in the data, e.g. due to measurement errors, into consideration (e.g. Verboon, 1994; Verboon and Willenborg, 1995). Other efforts did take 'noise' in the data into consideration (e.g. Paass and Wauschkuhn, 1985; Fuller, 1993). Unfortunately, these attempts have not yet produced a satisfactory model for the re-identification risk per record.

Somewhat less ambitious is a model for the re-identification risk for an entire microdata set, i.e. the risk that an unspecified record from the microdata set is re-identified. Again the SDC-rules would be very simple when such a model would be available. Only the keys that should be checked and the maximum risk that the statistical office that is releasing a microdata set is willing to take should be prescribed. If the actual risk for the entire microdata set is higher than the maximum risk then appropriate SDC-measures should be taken. In this case, it is not clear, however, which records should be modified by these measures, because no model for the re-identification risk per record is available.

A model for the re-identification risk per microdata set has been proposed by Mokken et al. (198Ϛ, 1992). This model takes three probabilities into consideration. The first probability, f, is the probability that a randomly chosen person from the population has been selected in the sample. The second probability, f_a, is the probability that a specific researcher who has access to the microdata set knows the values of a randomly chosen person from the population with respect to a certain key K. The third probability, f_u, is the probability that a randomly chosen person from the population is unique in the population with respect to a certain key K. Under various assumptions, some of which are unrealistic (e.g. that no measurement errors have been made), an expression for the re-identification risk per set, D_R, is derived, namely

$$D_R = 1 - \exp(-Nff_af_u), \qquad (1)$$

where N is the population size.

To evaluate this expression it is necessary to calculate f, f_a and f_u. The sampling fraction f is easy to calculate, of course. The other two probabilities, f_a and f_u, are more difficult to calculate, however. Evaluating f_a seems very hard, because this probability depends on the specific researcher and his knowledge of the population. To estimate f_u a number of models have been proposed in the literature. Models to estimate the number of uniques, and hence the value of f_u, include the Poisson-gamma model (Bethlehem et al., 1989: Mokken et al., 1989; Willenborg et al., 1990; De Jonge, 1990) and the Poisson-lognormal model (Skinner and Holmes, 1992; Hoogland, 1994). Because the results of these, and other, models are rather unreliable, and because it is very hard to evaluate f_a, the model by Mokken et al. cannot be used in practice to evaluate a re-identification risk per microdata set.

Another approach that is possibly of interest to gain an insight in the re-identification risk per record, although it does not provide a way to actually evaluate such a re-identification risk, is *fingerprinting*. The idea of this approach is that the records that the most likely ones to be re-identified are the records that are often unique on a low-dimensional key. An SDC-rule based on fingerprinting could be the following one: a record is considered unsafe, and hence may not be released unmodified, if it is unique on more than m k-dimensional keys. Such a rule cannot be applied easily, however, because the number of keys that have to be examined becomes astronomically large even for moderate values on k. For example, suppose that there are 50 identifying variables in a microdata set. Suppose furthermore that the values of m and k equal 10 and 6, respectively, i.e. a record is considered unsafe when it is unique on more than 10 keys consisting of at most 6 identifying variables. In this case, an upper bound for the number of combinations of variables that have to checked is about 16 million, equalling the number of ways to select 6 elements from a set of 50 elements. There are several ways to overcome this practical problem. Firstly, one can decide to consider only even lower-dimensional keys, say k equals 3 or 4. Secondly, fingerprinting is highly suited for parallel computing. In a distributed computing environment, e.g. a network of PC's, fingerprinting could be done very efficiently. However, further research is needed for fingerprinting to be effectively applied in practice.

So far in this section we have only discussed methods to evaluate the re-identification risk, either per record or per microdata set. There are several other issues that deserve further investigation. An important issue is a practical one, namely determining the local suppressions automatically and optimally in so-called complex microdata, i.e. microdata with logical and statistical dependencies explicitly taken into account. When determining the local suppressions these dependencies should be taken into account. For example, when the value of the variable 'Number of children you have given birth to' in a certain record equals '2', then local suppression of the value of the variable 'Sex' in this record does not offer any protection against disclosure, because it is clear that this value is 'Female'. Such dependencies can be incorporated in the 0-1 integer programming formulations to determine the optimal local suppressions (Cf. De Waal and Willenborg, 1995b). Efficient algorithms to solve the resulting problems (to good approximation) remain to be found.

5. SDC FOR TABLES

Between SDC of microdata and tables there are many similarities. For instance, when trying to reduce the risk of disclosure one usually starts by modifying the identifying variables. In case of microdata one collapses the categories of an identifying variable, in case of tabular data one collapses two columns or rows of the table. After the global modifications have been made local modifications must be made. In case of microdata values of identifying variables in some records can be changed to 'missing', in case of tabular data values of sensitive cells can be changed to 'missing'. This example also illustrates a difference between SDC for microdata and SDC for tabular data: in case of microdata we locally suppress values of identifying variables, whereas in case of tabular data we suppress values of sensitive data. In this section we examine such similarities and differences between SDC for microdata and tabular data.

First of all note that in the literature on SDC for tables it is generally assumed that the tables that are published are based on an observation of the entire population. The disclosure problem of tabular data in case only a sample of the population is observed is hardly discussed. In the sequel we also assume that the tables are based on an observation of the entire population.

After some columns and/or rows have been collapsed it is necessary to make some local modifications. A well-known technique to modify data in a table in order to safeguard this table against disclosure is cell suppression, which can be compared to local suppression in case of microdata. To apply suppression in tables the cells that contain sensitive information have to be determined. The usual way to determine whether a cell is sensitive is by means of a dominance rule. A dominance rule states that if the values of the data of a certain number of respondents, say 3, constitute more than a certain percentage, say 75%, of the total value of the cell, then this cell is sensitive. The main idea on which this approach is based is that when a cell is dominated by the contributions of a few respondents, then these contributions can be estimated rather accurately. For instance, if there is only one respondent then his contribution can be disclosed exactly. When there are exactly two respondents then each of these respondents can disclose the contribution of the other, and when the value of a cell is dominated by the contributions of two respondents, then each of these respondents is able to estimate the value of the contribution of the other one accurately. In general, if there are *k* respondents then *k-1* of them, after pooling their information, can disclose information about the value of the data of the remaining respondent. For small *k*, say, 2, 3 and 4, this poses a problem.

The sensitive cells in tables can be compared to the unsafe combinations in case of microdata. Like the unsafe combinations in case of microdata, sensitive cells have to be protected by suppressing, recoding or perturbing their values. We first consider suppression of sensitive cells.

The suppression of a cell because the contents of this cell is considered sensitive according to some sensitivity criterion, e.g. a dominance rule, is called *primary suppression*. Primary suppression alone is generally not sufficient to obtain a table which is safe for release. In a table the marginal totals are often given as well as the values of the cells. A cell which has been suppressed can then be computed by means of the marginal totals. Therefore, other cells have to be suppressed in order to avoid this possibility. This is called *secondary suppression.*

Like local suppression in case of microdata, secondary suppression in tables should be done in such a way that the information loss is minimised. Usually, weights are assigned to the cells in a table. The information loss due to suppression of a cell is then given by the corresponding weight. There are several possibilities to specify the weights. For instance, the weight of a cell can be chosen equal to the number of respondents in this cell. In this case, one aims at minimising the number of respondents whose data are suppressed in the table. Alternatively, the weight of a cell can be chosen equal to the cell value. In this case, one aims to minimise the total value of the data which are suppressed. Selecting 'good' weights is a matter of subjective considerations.

Secondary suppression causes other problems as well. Although it might be impossible to compute the exact values of suppressed cells in a table after secondary suppression, it is still possible to compute the ranges in which the values of the cells lie, when the marginal totals of the table are given and it is for instance known that the values of the cells are all nonnegative (Cf. Geurts, 1992). If these ranges are small for the sensitive cells, then an attacker is able to obtain good estimates for the values in the suppressed cells. Therefore, secondary suppression must be carried out in such a way that the ranges in which the values of the suppressed cells lie are not too small. Tables with marginal totals can be compared to complex microdata, because in both cases dependencies between (cell) values should be taken into account.

Another well-known technique to protect sensitive cells in a table against disclosure is *rounding*. The most interesting way of rounding is controlled rounding (Cf. Fellegi, 1975; Cox, 1987). The main advantage of controlled rounding compared to conventional rounding and random rounding is that the additivity of the tables is preserved, i.e. after rounding the rows and columns still add up to their rounded marginal totals. A slight disadvantage of controlled rounding is that a cell value is not necessarily rounded to its nearest integer multiple of the rounding base *b*, but rather to one of its two nearest integer multiples of *b*. Controlled rounding for two-dimensional tables does not provide serious problems any more. Note that rounding is a 'stylised' way of perturbing the data, i.e. adding noise to the data. Hence, to some extent rounding in tables is comparable to perturbation in a microdata set.

The protection offered by rounding to the sensitive cells should be approximately the same as when suppression would have been applied. In the case of suppression the range in which the value of a sensitive cell must lie should be sufficiently wide, say the width of this range should be at least *p*% of the cell value. The range of ambiguity offered by rounding should then also be at least *p*% of the value of a sensitive cell. From this criterion a value for the rounding base *b* can be derived for a given value *p*.

Three- and higher-dimensional tables and linked tables, i.e. tables with common variables obtained from the same base file, pose a lot of theoretical problems (Cf. De Vries, 1993). The theory for these kinds of tables is much more difficult than for ordinary two-dimensional tables. An interesting similarity between microdata and linked tables is the following. Suppose that one wants to protect a set of linked tables by recoding the variables. Suppose furthermore that one wants to use the same categorisation for each variable in each table where this variable occurs. The aim is to protect the linked tables in such a way that the information loss is minimised. This problem is similar to the so-called global recoding problem for a microdata set, i.e. the problem of applying only global recodings in such a way that the resulting data set is safe while the information loss is minimised. A solution for the global recoding problem for a microdata set implies a solution to the recoding problem for linked tables and vice versa.

For secondary suppression in three- and higher-dimensional tables some heuristic algorithms are available, but much work still has to be done in order to perfect these algorithms. This is another subject that attracts the attention of Statistics Netherlands. Controlled rounding of higher-dimensional tables is a difficult problem. In some cases the problem is impossible to solve (Cf. Cox, 1987). In these cases it is necessary to relax the conditions of controlled rounding. For instance, instead of demanding that each value is rounded to one of its two nearest integer multiples of the rounding base, one could specify a window of values in which the rounded value of cell should lie. Some heuristics to deal with three-way tables have been developed (Cf. Fagan et al., 1988; Kelly, 1990; Kelly et al., 1990). For four- and higher-dimensional tables satisfactory heuristics are hard to find. Another interesting problem is controlled rounding for linked tables. It is not possible to round each of these tables separately, because the same marginal totals occur in several tables.

6. DISCUSSION

As a consequence of the increased demand for detailed information Statistics Netherlands has disseminated a considerable number of microdata sets in recent years. The SDC-rules for these microdata sets were, and still are, based on intuitive reasoning and still lack a solid theoretical framework. The main idea of these rules is that the population frequencies of certain combinations of values of identifying variables have to be checked. The population frequency of such a combination should be sufficiently high, otherwise SDC-measures should be taken.

After the laborious process of developing the SDC-rules it was soon realised that applying them in practice is a nontrivial exercise. Instead it turned out that many methodological problems had to be solved. Some of these problems, such as determining the local suppressions automatically and optimally, have been solved (more or less) by now. Others, such as determining the global recodings automatically and optimally still remain to be solved. It was also realised that in order to apply the SDC-rules in practice a software package, ARGUS, should be developed. Without such a package the application of the SDC-rules would be very time-consuming and error-prone.

Although Statistics Netherlands has developed SDC-rules for its microdata sets, this does not imply that it is time to relax. In fact, we are somewhat dissatisfied with our rules. Better rules could be deduced if a good model for the re-identification risk per record would be available. Unfortunately, such models do not seem to be available at the moment.

Fortunately for researchers in the field of SDC, but unfortunately for statistical offices trying to protect their microdata sets, SDC for microdata sets offers many possibilities for future research. In the long run a model for the re-identification risk per record should be developed. Until that time, the present SDC-rules and techniques should be refined. For instance, global recodings should be automated and optimised, local suppressions (and global recodings) in complex microdata should be automated and optimised, and sampling weights should be protected efficiently. Only by continued research efforts the present and future challenges of SDC for microdata can be met adequately.

REFERENCES

Barnes, G. , 1995, Local perturbation. Report, Statistics Netherlands, Voorburg.

Bethlehem, J.A., W.J. Keller and J. Pannekoek, 1989, Disclosure control of microdata. Journal of the American Statistical Association, Vol. 85, no. 409, 38-45.

Cox, L.H., 1987, A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association, Vol. 82, 520-524.

De Jonge, G., 1990, The estimation of population unicity from microdata files (in Dutch). Report, Statistics Netherlands, Voorburg.

De Vries, R.E., 1993, Disclosure control of tabular data using subtables. Report, Statistics Netherlands, Voorburg.

De Waal, A.G. and A.J. Pieters, 1995, ARGUS user's guide. Report, Statistics Netherlands, Voorburg.

De Waal, A.G. and L.C.R.J. Willenborg, 1994a, Minimizing the number of local suppressions in a microdata set. Report, Statistics Netherlands, Voorburg.

De Waal, A.G. and L.C.R.J. Willenborg, 1994b, Development of ARGUS: past, present and future. Report, Statistics Netherlands, Voorburg.

De Waal, A.G. and L.C.R.J. Willenborg, 1995a, Statistical disclosure control and sampling weights. Report, Statistics Netherlands, Voorburg.

De Waal, A.G. and L.C.R.J. Willenborg, 1995b, Local suppression in statistical disclosure control and data editing. Report, Statistics Netherlands, Voorburg.

De Waal, A.G. and L.C.R.J. Willenborg, 1995c, Optimum global recoding and local suppression. Report, Statistics Netherlands, Voorburg.

Fagan, J., B. Greenberg and B. Hemming, 1988, Controlled rounding of three-dimensional tables, Statistical Research Division Report Series, Bureau of the Census, Washington DC.

Fellegi, I.P., 1975, Controlled random rounding. Survey Methodology, Vol. 1, 123-133.

Fuller, W.A., 1993, Masking procedures for microdata disclosure limitation. Journal of Official Statistics, Vol. 9, no. 2, 383-406.

Hoogland, J., 1994, Protecting microdata sets against statistical disclosure by means of compound Poisson distributions (in Dutch), Report, Statistics Netherlands, Voorburg.

Keller, W.J. and J.G. Bethlehem, 1992, Disclosure protection of microdata: problems and solutions. Statistica Neerlandica, Vol. 46, no. 1, 33-48.

Keller, W.J. and L.C.R.J. Willenborg, 1993, Microdata release policy of the Netherlands CBS. Proceedings of the International Seminar on Statistical Confidentiality, Dublin.

Kelly, J.P., 1990, Confidentiality protection in two- and three-dimensional tables. Ph.D. thesis, University of Maryland, College Park, Maryland.

Kelly, J.P., B.L. Golden and A.A. Assad, 1990, Controlled rounding of tabular data. Operations Research, Vol. 38, 760-772.

Mokken, R.J., J. Pannekoek and L.C.R.J. Willenborg, 1989, Microdata and disclosure risks. CBS Select 5, Statistical Essays, Staatsuitgeverij (The Hague), 181-200.

Mokken, R.J., P. Kooiman, J. Pannekoek and L.C.R.J. Willenborg, 1992, Disclosure risks for microdata. Statistica Neerlandica, Vol. 46, no. 1, 49-67.

Paass, G. and U. Wauschkuhn, 1985, Data access, data protection and anonymisation - analysis potential and identifiability of anonymised individual data (in German). Gesellschaft für Mathematik und Datenverarbeitung, Oldenbourg-Verlag, Munich.

Pannekoek, J., 1992, Disclosure control of extreme values of continuous identifiers (in Dutch). Report, Statistics Netherlands, Voorburg.

Pannekoek, J. and A.G. de Waal, 1995, Synthetic and combined estimators in statistical disclosure control. Report, Statistics Netherlands, Voorburg.

Pieters, A.J. and A.G. De Waal, 1995, A demonstration of ARGUS. Report, Statistics Netherlands, Voorburg.

Skinner, C.J. and D.J. Holmes, 1992, Modelling population uniqueness. Proceedings of the International Seminar on Statistical Confidentiality, Dublin.

Van Gelderen, R., 1995, ARGUS: Statistical disclosure control of survey data. Report, Statistics Netherlands, Voorburg.

Verboon, P., 1994, Some ideas for a masking measure for statistical disclosure control. Report, Statistics Netherlands, Voorburg.

Verboon, P. and L.C.R.J. Willenborg, 1995, Comparing two methods for recovering population uniques in a sample. Report, Statistics Netherlands, Voorburg.

Willenborg, L.C.R.J., R.J. Mokken and J. Pannekoek, 1990, Microdata and disclosure risks. Proceedings of the 1990 Annual Research Conference, Bureau of the Census, Washington DC, 167-180.

Willenborg, L.C.R.J., A.G. De Waal, R.E. De Vries and C.A.W. Citteur, forthcoming, Statistical disclosure control in practice. Springer-Verlag, New York.